Framework for the definition of significant properties

Work Package:	3.3
Document Type:	Public
Author:	Gareth Knight
Contributors:	Stephen Grace, Lynne Montague
Evaluators:	Ian Hodges, Adrian Brown, Paul Wheatley
Draft/Version:	V1
Date of edition completed:	05/02/2008
Change History:	
Version 1 – 14/02/2008	Public version. Some editing of document text, based on feedback from Paul Wheatley, Lynne Montague & Adrian Brown
Draft 4 - 05/02/2008	Revision based on feedback by project team
Draft 3 - 14/01/2008	Re-ordering of document
Draft 2 – 12/1/2008	Creation of preliminary list of significant properties for audio, structured text and email
Draft 1 - 10/10/2008	First draft

Contents

Abstract

Significant properties are those aspects of a digital record that must be preserved over time in order for it to remain accessible and meaningful. The development of a formal, or canonical[a] method to define significant properties requires some understanding of the digital record, the different types of properties that may be encountered and the appropriate method for their expression. In this report, we will outline a methodology for the identification and description of significant properties contained by a digital resource. We also provide a generic template that may prove useful for the description of a wide range of digital resources.

1. Introduction

The curation of digital data has been an area of intense research during the previous 10-20 years. The use of computing technology can present many problems when considered in the long-term: hardware and software is often replaced with new products that offer only limited backwards compatibility leading to digital information that can no longer be read in its native format and which requires the use of decoders to be understood by a user. In combination, these factors may create situations in which digital information continues to have intellectual value, but becomes increasingly difficult to access as a result of technology obsolescence. For anyone who wishes to continue to access the digital information, several strategies are available. These include:

1. **Technology preservation:** maintaining the technology on which the original software was executed.
2. **Technology emulation:** recreating the original operating environment on a new platform.
3. **Software recompilation:** converting the original software to a new platform.
4. **Specification re-creation:** re-creation of the original software specification in different software.
5. **Content conversion:** converting digital information to different encoding format and software applications.

Each one of these strategies is practical and useful to perform in different circumstances. However, the level of investment required, in terms of the time and knowledge required to implement the strategy, is variable. There are also potential dangers in taking each strategy. Strategies 2-5 imply some form of reinterpretation, which introduces the potential risk that the process will not be performed correctly and that some information may be altered or excluded.

To authenticate that the required information is complete and unchanged in comparison to the original Record, some form of validation is required. In simple cases, a manual inspection of converted data is sufficient to identify obvious errors that have occurred. However, the approach is likely to be impractically time-consuming for the review of a large number of files. Instead, it is preferable to develop a machine-processable assessment criteria that may be used by automated tools to examine each resource, compare it to an original and validate that the significant properties are unchanged. In this report we describe a methodology for the definition and evaluation of significant properties contained in digital records. The methodology outlines factors that must be considered when identifying properties that are essential and makes recommendations for evaluating their relative value. The approach taken is illustrated through the analysis of four file types that contain different types of information.

1.2. Definition

The concept of significant properties has been a focus of analysis and reference by several projects during the previous 10 years. The OAIS (Open Archival Information System) Reference Model (CCSDS, 2002) is the most influential document for understanding data management requirements, indicating the workflow activities that must be performed to

maintain data. Although the reference model does not explicitly refer to 'significant properties', the concept may be identified in the foundation of which the model was built – the conversion of the Information Object contained in the Submission Information Package (SIP) into a form appropriate for archival or dissemination purposes. In the context and terminology of OAIS, significant properties are the characteristics of the Information Object, encoded in a digital object that must be reproduced, even if there are changes to the hardware and software in which the Information is created and managed. The JISC CEDARS[b] and CAMILEON[c] Projects, funded during the late 1990s explored several concepts introduced by the OAIS and provided an explicit link between the conversion of digital data and significant properties. They indicated that the significant properties are closely linked with the need to maintain the authenticity (the establishment of its purpose and the processes through which it was created and maintained) and integrity (that it has not been changed or corrupted in a manner that has caused the original meaning to be lost[d]) of the Record.

An implicit assumption in the use of terminology, such as 'significant' and 'essential' is the recognition that an assessment criteria is required against which the relative value of each property may be assessed. The assessment criteria must establish a set of requirements or an objective that each property has to fulfill. In recent years, several authors have attempted to provide an appropriate definition of significant properties that encapsulates their function and requirements. In an earlier work package, Wilson (2007) summarized the different approaches and used them as the basis to create a definition of significant properties, as they apply to the archival community that encapsulates three functions:

> "the characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects."[e]

A notable insight that the OAIS Reference Model provides is the recognition that the level of significance attributed to each property is subjective and may change over time as it is tailored to the needs and capabilities of the institution and the activities it performs. For example, the requirements of the Designated Community may become increasingly sophisticated over time or the knowledge base of the Designated Community required to understand the data may decrease. Alternatively, the granularity at which the significant properties are defined may also change, according to the capabilities of the software tools. In each scenario, the type of properties and the information required to understand them will change.

1.3. Methods for the identification of significant properties

To determine the significant properties of a digital Record, a consistent, formal method of identifying the important aspects is required. The National Archives of Australia (2002) has developed a 'Performance Model', which has been adopted by the InSPECT Project. The Performance model establishes the concept of the 'essence' of a digital record that contains the "characteristics that must be preserved for the record to maintain its meaning over time." The principle of the model is that the re-creation of Information Content relies on an interaction between the underlying data and the technology used to produce an output. The Source is the basic Data Object i.e. the bits that constitute the text, still images or moving images file that must be interpreted; Process indicates the software used to interpret the Data Object and extract the information; and Performance indicates the rendering of the Information Content in a format that is understandable to the user. To illustrate its application to different scenarios, figure 2 indicates how it may be applied to an OAIS-based scenario, in which 'Representation Information' – the information required to access and interpret the Intellectual Content - is used to interpret a set of raw data, interpret its content and render it as an Information Object in a format that may be understood by the user.

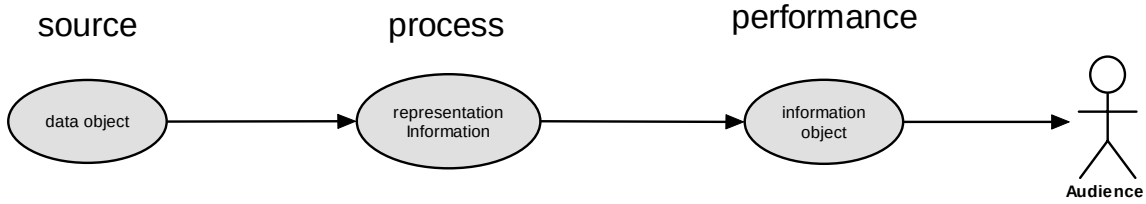
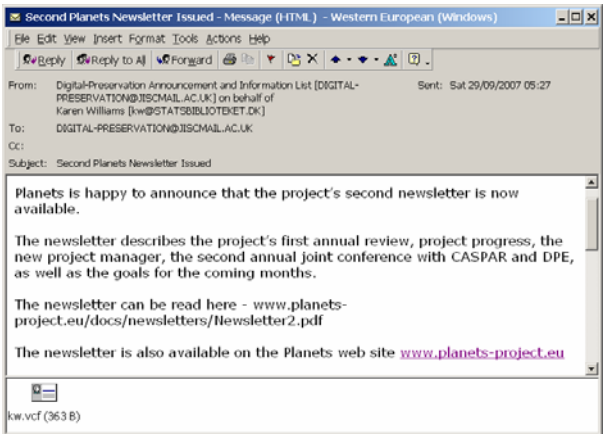
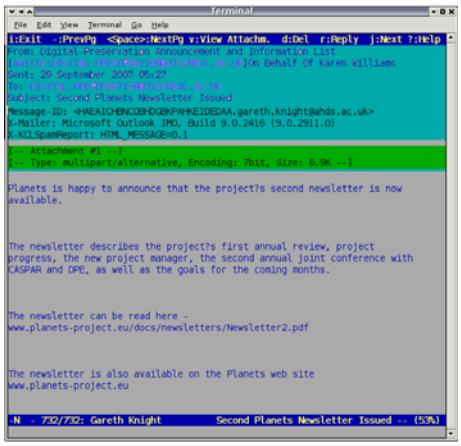


Figure 2: The application of the Performance Model to the re-creation of an OAIS Data Object

In recognition of the changing technological environment and the requirements of the Designated Community, the process required to re-interpret the source may vary between computer platforms and change over time. To illustrate the former, an email represents one example of a Source that may be 'performed' to a user. Its significant properties relate to the content contained in the message body, contextual information regarding the sender, recipient and subject, as well as details of any attachments that may be provided. However, the appearance is, in many cases, not considered to be important and is subjective to interpretation by different software applications to create the performance. Microsoft Outlook operating in Microsoft Windows displays the email on a white background with a grey border to indicate header information; Mutt operating in a Linux terminal displays the text in blue on a grey background.



Microsoft Outlook running on MS Windows 2000



Mutt running in a Linux terminal

Figure 3: Two different interpretations of an email

Figure 3 provides a simple example of how the appearance of a simple type of Source will differ between software applications. It is possible to envisage scenarios in which the Performance of complex file types, such as 3D models, differs significantly between software applications to such an extent that the meaning is changed and authenticity is lost.

1.4. A framework for recording significant properties

The development of a canonical list of significant properties is considered to be a management activity that may be used to guide the assessment of format conversion and emulation activities, allowing the assessor to identify if any information has been lost and to measure its relative value to the re-creation of the Performance in its entirety. An aspect of the work of the InSPECT Project has been the creation of a framework that may be used to catalogue the significant properties associated with a digital Record. The creation of a significant properties framework fulfils several purposes. It enables the institution to:

1) Analyse and catalogue the significant properties of a digital Record;
2) Review significant properties associated with an existing digital Record;
3) Assess the relative value of the property for the re-creation of the Record;
4) Quantitatively measure the value associated with the property;
5) Validate that the value associated with the property is correct.

Handling guidelines of the type described exist in many institutions. However, they are commonly written as procedural lists for a Curator to perform, separate from the Record itself. By storing the significant property information as metadata with the Record itself, the representation information may be transferred between digital repositories.

The project team sought to define a common set of information elements, similar to the Dublin Core metadata standard that may be used by an institution to identify the properties considered to be important and indicate the quality thresholds that must be met. The

framework, illustrated in figure 4, may be incorporated into appropriate format registries, such as PRONOM or the GDFR and/or metadata schemes, such as that associated with the PREMIS Data Dictionary as appropriate.

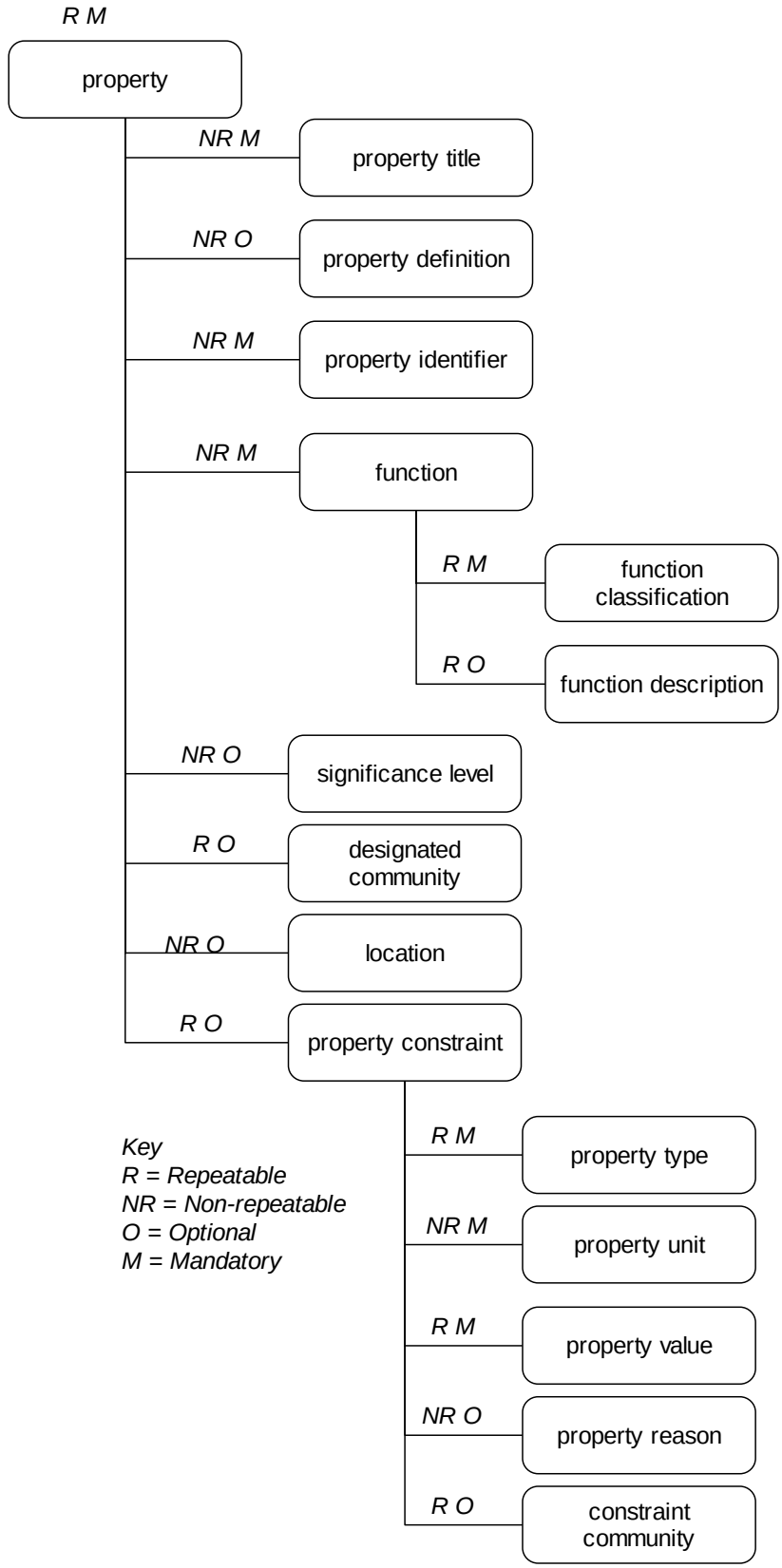


Figure 4: a framework for the description of significant properties

The containers, sub-containers and elements illustrated above perform different functions necessary to identify, describe and measure the significant properties of a Record. The following definitions are used:

- *propertyTitle:* The title of the property that indicates its purpose. The archive should maintain consistency across file types by using consistent terminology for properties that perform the same function. The same propertyTitle must not be used for two properties that perform different purposes.

- *propertyDefinition:* A formal statement that describes the purpose of the property. The definition provides a human-readable description of the property. It should be stored by an appropriate service provider, such as The National Archives' PRONOM and is not intended to be stored in the Record metadata itself (see Table 4 for examples)

- *Property Identifier:* A machine-processable identifier to categorise each property.

- *function*
    - *functionClassification:* A controlled vocabulary that indicates the high-level function that the property performs in the Record. For example, Content, Context, Structure, Rendering and Behaviour (see page 14)

    - functionDescription: A free text description of the function that the property performs (see Table 3 for examples)

- *significanceLevel:* An assessment of the significance of the property to the re-creation of the Record (see page 12)

- designatedCommunity: The importance attributed to a property may differ between designated communities. The designatedCommunity value allows the archive to declare the properties that are important to specific user types. Possible examples of two designated communities are 'archive' for institutions performing preservation and 'dissemination' for academics and other users. By leaving the Designated Community value blank, the archive declares that the property is, as far as they are aware, important for all user communities.

- *location:* The field should be used to indicate the layer at which the property is applicable. A property may be associated with a Record or Component (see below).

- *propertyConstraints*

    - propertyType: An indicator of the type of constraint placed on the value of the property. Three constraints are currently recognized:
        - equality: the property stored in the Record must be equal to one or more values stored in the metadata.
        - minimum: if a numeric measurement is used, minimum indicates the lowest numeric value that is allowed. The minimum and maximum measurement types must be used in combination.
        - maximum: if a numeric measurement is used, maximum indicates the highest number value that is allowed. For example, the highest sampling rate of an audio recording.

    - propertyUnit: The unit in which the value is measured. E.g. hertz, no. of characters.

    - propertyValue: The measured value of the property or the location in the technical metadata where it may be located. The meaning of the propertyValue will differ according to the propertyType. For example, a value of '96000' may indicate the highest sample frequency value that is allowed; if the propertyType is 'equality' the propertyValue may contain an exact value against which subsequent Performances must be measured (e.g. the subject

line of an email), a measured value (e.g. number of characters, colour value), or some other measurement. A blank value may be entered if the archive recognizes that the property is important, but does not possess the appropriate software to measure it. However, this use is discouraged. If the archive stores the property measurement in other technical metadata, the propertyValue may be used to indicate the location (e.g. its location in a METS document).

- propertyReason: A free text field that may be used to explain the rationale for the constraint. Although no constraints are placed on the type of information include in the field, institutions are encouraged to take a consistent approach to descriptions.

- Communityconstraint: The communityConstraint value allows the archive to declare the acceptable property constraints that may be tailored for different user types. A constraint may be applied to one or more specified groups in the Designated Community. Possible examples of two designated communities are 'archive' for institutions performing preservation and 'dissemination' for academics and other users.

The significant properties framework has been applied to a preliminary list of the properties that are considered important for the curation of the four types of digital file being analysed in the project (consult section 3 for practical examples and the appendix for an assessment template). However, further work must be performed to test and refine the framework on the significant properties of other types of digital record.

A second associated activity is to identify the level of granularity at which the significant properties of a digital Record may be described, as indicated by the 'location' value. We may begin to understand the level of granularity at which significant properties may be assessed and the relationship between entities by reviewing the data models that are in widespread use. It is beneficial to consider the conceptual data model that is to be used prior to the classification of significant properties.

Several conceptual models exist that have been developed to fulfill a wide range of different scenarios. Common data modeling techniques that may be familiar to the reader include the FRBR[f], ABC[g] and PREMIS[h] data models. The TNA conceptual data model was produced by The National Archives for the Seamless Flow programme and has subsequently been adapted by the PLANETS Project[i]. The detail provided by each data model differs, ranging from the conceptual to the technical properties of a Record. However, they are broadly compatible. Table 1 indicates the key terms in each data model and the relationship between the various entities.

TNA	**PREMIS**	**FRBR**
-	-	Work
Deliverable Unit	Intellectual Entity	Expression
	Representation	
Manifestation	-	Manifestation
-	Object	Item
File	File	-
Bitstream	Bitstream	-
-	Filestream	-

Table 1: Key terms in the TNA, PREMIS and FRBR data models

For the purpose of this document, the InSPECT Project team adopted a simplified version of the TNA Seamless Flow data model, defining the Performance as a compound of many types of information (e.g. text, images, sounds, etc.) consisting of intellectual or technical components, which serve as logical sub-groups of information, e.g. a shape in a 3D model, an email that consists of context information, etc. Each of these represents a part of the whole

that must be maintained and are likely to possess technical properties. Figure 5 illustrates the relationship between the Record and Component.

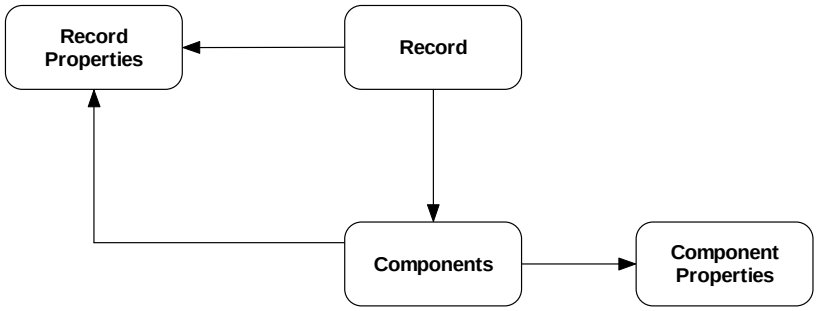


Figure 5: A conceptual model for identifying the key Components of a Record

The data model developed for the description of significant properties is relatively simple in its design. However, it offers some degree of flexibility to describe a range of properties at different levels of granularity. Further testing is necessary to produce a concept model that is appropriate to a wider range of digital Records. Discussion is ongoing between the four JISC-funded significant properties projects on the creation of a unified data model that will allow the definition of the properties of each file type at an appropriate level.

2. Methodology

The definition of properties that should be considered significant for different classes of record content is not a simple task that can be automated based on a set of universal rules. A set of rules defined for one category of resource may prove to be too restrictive when applied to unusual variations, or inappropriate for other file-types[1]. Instead, the InSPECT Project team has sought to take an epistemological approach, by considering the intended purpose of the Information Content during evaluation.

The InSPECT project investigated the significant properties associated with four file types - raster images, digital audio, structured text and emails. The project team read through numerous format specifications and recorded the key components of each type. On completion of a list of record properties, the project team developed and tested an Assessment template (see appendix 1) that was used to evaluate each property. The assessment template required the assessor to record each property, document its use and consider the most effective method of classifying the property according to the function it performs and its importance to the intellectual content in its entirety. This has proven an effective analysis method, identifying some properties that are common to certain creation methods and content types and others that are used for specific purposes.

The methodology developed for the project may be separated into four key stages:

1) Definition of the intellectual Components of a Record that the assessor wishes to maintain
2) Identification of the technical properties of each Component that are required to recreate it
3) Classification of the function performed by each property and an assessment of its value
4) Measurement of the technical and intellectual properties of the digital Record, as well as consideration of the requirements of the institution and Designated Community.

Each one of these tasks has been separated under an appropriate heading that describes the underlying principle for understanding the tasks. Practical examples of the methodology are located in section 3, as applied to structured text, e-mails, raster images and sound files.

2.1. Definition of the intellectual Components

The definition of the Components that serve as the basis for the re-creation of the Record is, to some extent, an intellectual exercise and, as a result, may be overlooked in favour of technically-oriented approaches to the extraction of Record properties in which the decisions are made for the user. The identification of something as significant requires a pre-defined criterion against which it is measured. The classification of certain properties as significant cannot be performed in isolation from the larger purpose of the digital record. It requires consideration of four factors:

1) The standards with which the institution that is maintaining the Record is required to comply
2) The requirements of the Designated Community
3) The Component's contribution to the re-creation of the Record as a 'whole'
4) The capabilities of the tools to maintain each property.

In combination, the four criteria encapsulate a combination of objective and subjective decisions.

The first factor is the simplest to identify in advance. National institutions, such as The National Archives and The British Library have a remit to comply with the BS ISO 15489 standard, which regards a reliable record as one "*whose contents can be trusted as a full and*

[1] For instance, the significant properties of an structured text object, such as an email and an unstructured text object, such as a letter differ, although they share similar technical characteristics.

accurate representation of the transactions, activities or facts to which they attest"[j]. In many circumstances, it is relatively simple to identify the properties of a digital Record that fulfil the criteria of establishing its authenticity. For example, an email may be considered authentic if the message body is maintained, in addition to context information that establishes the sender, recipient, sent and received date. Similarly, the authenticity of a digital audio recording may be established by maintaining the audio recording and associated metadata.

The identification of digital properties that are significant in relation to the second and third factors is more difficult, requiring some degree of subjective judgement. An assessment may be influenced by the subjective evaluation of the informative potential to the Designated Community, which introduces an element of uncertainty. Record properties can have different meanings for different purposes or scientific disciplines. Scientific disciplines or theories can have different foci or different epistemological interests[k]. Staff with particular expertise in a subject of research, such as library or information science staff, may have the required depth of subject knowledge to make a decision on the needs and requirements of their Designated Community. In the absence of an objective calculation of significance for different properties in each context, an epistemological approach may be taken by defining the common functions that different types of digital Record must perform. Table 2 indicates the intellectual Components that are common to various types of relatively simple digital records[2], as categorised by the data type and intended purpose.

File Type	Purpose	Component	Comments
2D Still Image (raster)	Photograph	Pixels	
		Description	E.g. creator, GPS location, etc.
	Page Scan	Image	
		Description	
		Text	An image may be scanned of a page of text.
2D Still Image (vector)	2D graphic		Further details of the intellectual Components of a Record may be found in the study on Significant Properties of Vector Graphics
		Shapes	
		Text	
		Other 2D Components	
Computer Aided Design	3D object	Shapes	
		Other 3D Components	
Audio	Sound recording	Channel1 [left channel]	A distinction may be made between the intellectual value of sounds stored in different channels, e.g. a channel may

[2] For the purpose of simplicity, at this stage the Presentational structured text document excludes details of the relationship between sub-components and the list of components for email excludes the possibility that it will have file attachments or other relationships. These considerations are explored in a subsequent section.

			contain audio important to create a surround-sound effect or, may contain unwanted noise.
	Sound recording	Channel2 [right channel]	
	Sound recording	Description	
Unstructured Text	Text document	Body	
Structured Text	Email	Creator	
		Sender	
		Primary Recipients	
		Secondary Recipients	
		Sent Date	
		Received Date	
		Keywords	
		Message-Body	
Structured Text (Presentation)	Web Page	Paragraphs	
		Lists	
		Tables	

Table 1: taxonomy of Intellectual Components that may be located in different types of digital record

Finally, the practical capabilities of the software tools used to identify, distinguish between and separate intellectual components should also be considered. For example, an assessment of the value of different properties in a Computer Aided Design (CAD) diagram is unnecessary, if it is not possible to separate and remove the superfluous aspects. However, the assessor should be wary to avoid making value judgments on the basis of technology limitations, which are likely to be improved and enhanced in the future.

By separating the intellectual Components of a Record, the assessor can begin to consider the preservation requirements of each one in turn and identify properties that may be superfluous to its performance. The review of intellectual Components that must be preserved may be performed in relation to the preservation policy of the digital repository and/or the requirements of the Designated Community.

2.2. Definition and valuation of the technical properties

A second activity, following on from the identification of the intellectual Components that need to be preserved is the identification of the technical properties on which they are based. The number and type of properties that are significant to the re-creation of each intellectual Component are diverse. The property may directly contribute to the re-creation of the intellectual Component, or indirectly through being required by another property.

The evaluation of the contribution of each property to the re-creation of a Component and its contribution to the record as a whole may be assed via two subjective perspectives: an assessor may take a 'risk adverse'[I] approach that considers the property to be essential if it is present in the original Record, or an economic approach that will reduce the complexity of the conversion process and, potentially reduce the amount of funding and time that must be allocated to it. The latter is considered to be the most effective method and underpins the InSPECT Project's approach to the evaluation of significant properties. To identify the technical properties in a Record that MUST be maintained three factors should be considered:

- **Function:** what is the function that the property performs in relation to the intellectual Component or the Record in its entirety?
- **Uniqueness:** Do other properties exist that perform a similar function and could be substituted?

- **Robustness:** What effect will it have on the re-creation of the Performance if the property is damaged?

Each of these factors may be extended and subsequent questions considered. Methods of addressing the first and second question should be performed in conjunction with the development of a Property classification taxonomy further described in 2.3. The third factor regarding the robustness should be considered in relation to section 2.4 on property measurement.

During the analysis of the four file types, it was recognised that properties fulfil two different purposes, providing information on the Intellectual Component or the environment required to reproduce it. To distinguish between properties defined for each purpose, the following terms have been defined:

- **Logical properties:** Properties contained in a Record that provide some explanation of the characteristics of the intellectual Component itself. For example, the duration of an audio recording, the start and end point of a line in a vector diagram. Logical properties are likely to be well defined and are essential to the correct reproduction of the Intellectual Components contained in the Performance.
- **Environment properties:** Properties contained in a Record that indicate the ideal environment in which the Intellectual Component may be reproduced. For example, the bit depth and sampling rate of an audio recording may be set to an appropriately high value that is considered to be suitability safe to reproduce the audio quality, although the actual audio recording may have been recorded at a much lower quality and gain no benefit from the higher value. The values assigned to environment properties may be considered more subjective than Logical properties and may be changed for different performances of the Intellectual Content[3]. During the analysis of the four content types it was recognised that those properties that contain environment information frequently appear in the Rendering and Behaviour categories, may occur in Structure category and are not present in the Content and Context categories[4].

An institution, particularly one with an obligation to maintain Records is likely to be risk adverse and will require the exact values associated with each property for preservation. However, there may be circumstances in which a lower quality derivative may be created. In these circumstances, it is useful to identify the properties that cannot be altered and the properties that allow some variation.

Although many institutions are willing to preserve the significant properties of a Record, many institutions may find it difficult to perform the task. It can be a time-consuming process to deconstruct and evaluate each property; the assessor must have a good technical background in order to assess the value of the property, in terms of its contribution to the Record as a whole; and the format specification must be available and well documented. To evaluate the relative significance of a property through the use of a five-point scale, the InSPECT and Significant Properties of Vector Images projects[m], as well as early work by the SPELOS (Significant Properties of Learning Objects) and Significant Properties of Executables [n] projects have adopted a common set of performance indicators that may be assigned by an assessor.

Numeric Value	Summary	Description
10	Essential and unchanged	Removal or damage to the property is likely to result in the inability to use or reproduce the performance.
07-09	Essential. Some variation allowed.	The property should be maintained to recreate the Performance. However, the

[3] For example, the quality of a Master Record intended for preservation is likely to be encoded at a higher quality level in comparison to a distribution version.
[4] See section 2.3 for details of these categories which are assigned to each property.

		value assigned to the property may be changed to some degree, intentionally or unintentionally without significant effect on the re-creation of the performance.
04-06	Beneficial	The property is used in the Performance and may be maintained. However, other properties exist that perform the same or similar purpose.
01-03	Minor contribution	The removal of, or damage to, the property results in minor loss and does not contribute to significant loss to the Performance.
0	Not Applicable	The property is unimportant for the reproduction of the performance and does not contribute to the semantic understanding or use of the performance.

Table 2: Measurement of significance

A preliminary evaluation of the significance of different properties to the four object types is located in section 3. When applying the evaluation criteria to other object types, it is recommended that 0, 02, 05, 08, and 10 are used as normative values for the evaluation of each property. The values above and below these figures, within the specified ranges, may be used as the basis for additional weighting, specified by subjective assessment of the institution. Possible criteria for the addition of subjective weights may include the need to identify one or more properties that perform a similar function as being of greater or less importance. In these circumstances, the institution may consider assigning a '04' for properties that are less important in comparison to a second property with a value of 05.

2.3. Classification system for property functionality

To distinguish the properties that are essential from those that are superfluous, an assessor must have a defined understanding of the function performed by each property and its contribution to the whole. Previous work in the area, performed by Rothenberg & Bikson (1999)[o]; the InterPARES Project[p]; and the Digital Preservation Testbed (2003)[q] have recommended the creation of a classification system as a useful means to categorise information into one or more logical structures. One of the tasks assigned to the InSPECT project was the creation of a taxonomy that may be used to define and describe the function(s) of a broad range of Record properties. In the preliminary specification, it was identified that the taxonomy should describe the Logical and Environment functions indicated in section 2.2.

- Indicate the properties that are important to maintain the intellectual Components of a Record that contains a particular type of information.
- Define the function that each property performs in the reconstruction of the Record.
- Classify and identify properties that perform a similar function at an appropriate level of granularity.

In a previous work package, Wilson (2007) recommended that the five category terms and definitions (Context; Content; Structure; Appearance; and Behaviour) provided by the Digital Preservation Testbed (2003) were used as high-level categories. The suitability of these categories were subsequently assessed and extended to produce a canonical classification system for different types of file properties. During the analysis, several changes were made to the classification system, the most notable being the renaming of 'Appearance' to 'Rendering'. It was thought that Appearance implies the category may be applied to visual Components only. By using the term, 'Rendering' the category may be extended to incorporate the re-creation of non-visual Components.

For the purpose of analysis, the InSPECT project team has utilised a taxonomy based on these five categories:

1) **Content**
   Content is an abstract term to describe the expression of intellectual Work. In a digital environment, Content may describe text, still and moving images, audio, and other intellectual productions.
   Examples: logical properties: duration, character count.

2) **Context**
   Context may be applied to any information contained in the digital record that describes the environment in which the Content was created or that affect its intended meaning.
   Examples: Creator name, date of creation, description of the intellectual work, computer environment in which the Source was created (possibly).

3) **Rendering**
   The rendering category refers to any information that contributes to the re-creation of the performance. For example, it may be applied to a visual or audible Component.
   Examples: font type, colour and size, bit depth.

4) **Structure**
   Structure refers to any information that describes the relationship between two or more types of Content, as required to reconstruct the performance. It may be applied to the intrinsic or extrinsic relationships contained in the performance.
   Examples: E-mail attachments

5) **Behaviour**
   Behaviour is applicable to any information that describes the method in which the Content interacts with other stimuli. Stimuli may include the interaction of the user with the software, or the interaction with other sources of information, such as an external resource that affects the context, content, structure, or appearance of the resource. Behaviour is considered to be the most difficult characteristic to preserve – it is often tied to the capabilities of particular software applications and may be difficult to translate. It is also difficult to define all behavioural characteristics in a quantitative manner.
   Examples: Hyperlinks

These five categories may be supported by additional terms that indicate the Component of the Record to which it is applied in further detail. The ability of the assessor to define Components may differ, according to the file type (e.g. raster image, vector image, email, etc.), granularity of the analysis (e.g. the description of a property that contributes to the creation of a particular Component) and, for obvious reasons, the purpose of the property. To provide an example, Table 3 provides an example of the classification system that may be used to identify properties contained in an email header.

Record Type	Function Classification	Function description	Property Value
Email	Context	Creator	local-part
			domain-part
			domain-literal
			display name
		sender	local-part
			domain-part
			domain-literal
			display name
		Reply-to	local-part
			domain-part
			domain-literal
			display name
		Recipient(No.)	local-part
			domain-part
			domain-literal
			display name
		Sent-date	Date
		Received-date	Date
		keywords	Keyword1
			Keyword2
			Keyword3
			Keyword4

Table 3: A classification system for conceptual and technical properties

Section 3 of the document illustrates the significant properties that may be applied to different file types and indicates a classification for use.

2.4. Measurement of property values

A third and final stage of investigation is to define the method of measuring the properties of a Record. To apply an economic argument to the definition of significant properties: if it may be identified, it can be measured. This serves as a method to identify that, not only has the property been transferred when performing format conversion, but also that it has been transferred correctly and that the integrity of the Record has not been damaged. For this analysis, it is recognized that the properties of a Record may be measured using four methods:

1) **Identify:** Identifies the presence of a property in a Record. The property may be measured through a Boolean, indicating if the property is present or absent.
2) **Populated:** Identifies if a value is associated with a property, e.g. is the Creator field populated or empty?
3) **Measure:** Measures the conformance of the property to an expected norm e.g. the value is numeric and is within a pre-defined range; the value in the recreated Performance is an exact recreation.
4) **Validate:** Confirm that the property remains the same over subsequent manifestations of the same Record.

A key issue to consider is the sensitivity of the value(s) that is stored in the property: is it necessary to recreate the property exactly, or is it acceptable to allow some degree of variation? The answer to the question is likely to depend on the type of property being analysed, its function to the re-creation of the Record and the Designated Community. A logical property that refers to the intellectual content of the Record itself must remain the same. However, an environment property that controls the re-creation of the Performance may, to some extent be altered to suit the environment. In these circumstances, it is recommended that an Upper and Lower specification limit is developed that indicates the allowable deviation from the target value where a characteristic continues to be valid for the representation of the Record. To demonstrate, figure 1 specifies a hypothetical upper and lower limit for the sampling rate.

Figure 1: The acceptable upper and lower specification limits for a property

In this example, the 'target' sampling rate, as contained in the audio source is 48,000Hz. The tolerance specification indicates that a numeric value between 44100Hz and a hypothetical maximum of 96000Hz is acceptable to the digital archive. The digital archive has indicated 44100Hz as the lower specification that it will accept on the basis that further reduction in the sampling rate will cause a quality reduction.

The measurable difference between the lower specification, target and upper specification limit should be developed with some consideration of the property type and relative importance to the overall performance.

2.5. Methodology summary

The methodology described indicates the factors that must be considered to identify, classify and measure the property of a digital Record. Most notable, is the need to have a clear understanding of the key Components that perform a specific function that you wish to preserve. The following sections provide a preliminary list of the properties that are considered to be important to different types of digital object.

3. Initial analysis of file types

The authenticity of a digital resource is considered to be of great importance for the preservation of information. A limited number of core properties are common to all digital resources – all files stored on a file system will contain details of the creation date and last modified date. However, the majority of properties are likely to be unique to digital resources of a particular type. For the purpose of analysis, four file types were identified for assessment by the InSPECT project – raster images, digital audio, structured text, and e-mails. In the following section, we will apply the assessment template (see appendix) to the four file types and identify properties considered important to recreate the performance of the resource, as well as maintain its integrity and authenticity.

3.1. Structured Text

3.1.1. Definition

Structured text may be considered a catch-all term for a wide range of different types of content, encoded using a number of methods. It may be understood as electronic data that contains text, represented by alphabetic, numeric and punctuation characters, accompanied by information that indicates its description or appearance. The key characteristic that distinguishes structured and unstructured text is the presence of markup that provides additional information about the text. Structured text may be created for two purposes:

1. *Presentation* – Markup intended to describe the display of textual content. It may be used to infer the structure or layout of textual content, e.g. text rendered in bold or a large font may indicate a title or column heading and italicized information may indicate emphasis or particular display conventions, such as indicating the author of a work.

2. *Description* – Markup intended to indicate the semantic meaning of text, but not the method in which the information may be utilized. It is an exercise for the software application or researcher to decide on the method with which markup is handled. For example, software may extract text that is encased in a <creator> for use in the creation of a coversheet, or may attribute different display characteristics (bold, italics).

Presentation and descriptive markup languages separate information into logical structures. However, the principle for defining categories of information differ – presentation markup is primarily intended to affect the visual representation of a page (e.g. text emphasis, page layout); descriptive markup separates information categories into the appropriate semantic meaning. A digital Record may contain presentational markup,descriptive markup, or a combination of both.

Many presentation formats can be considered to be compound objects that are comprised of a primary Component and several associated secondary Components, e.g. images, sounds, etc. The Information Content contained in the compound object may be presented using a number of methods – through the primary Component in isolation; through a combination of the primary and one or more Secondary Component; or through the Secondary Component in isolation.

For the purpose of analysis, this report examines the requirements of structured text containing a mix of presentation & semantic markup. This report considers the preservation requirements of compound objects that consist of textual information (Primary Component), and a combination of textual and other information (Primary and Secondary Component). The third method of presenting the performance, as detailed above, may include a range of additional factors, dependent on the type of information contained in the Secondary Component, so is considered to be out of scope. This document will include some consideration of HTML and XHTML-based markup. It does not include a discussion of binary text documents that, although broadly similar, have other preservation requirements that must

be considered. It also excludes an analysis of structured text files that contain dynamic content that may change, based on interaction with the user.

3.1.2 Application of the Performance model to structured text

The central premise of the Performance model is the distinction between the raw, un-interpreted data, defined as the Source, and the interpretation of the data as a Performance. Although this is a useful metaphor, its application for structured text documents will vary, as distinguished by the content type and the rendering method. During the analysis it was recognized that, when applied to certain types of structured text (e.g. XML documents that do not possess associated instructions on the preferred method of recreation), the Performance Model metaphor is unhelpful unless a distinction between the Source and Performance can be made. Many types of structured text may be 'performed' using several methods. The purpose of our analysis is to describe the performance of structured text in a particular environment. It does not, and indeed cannot, describe every type of performance that can be made of structured text. To illustrate, an XML-encoded text may be presented to the user as an RSS feed, processed and converted to an audio stream, and/or represented in several XHTML-compliant web pages that contain different types of information (figure 6). If a theatre Performance metaphor is applied, it may be compared to the recreation of a script by one or more actors in different theatre environments.

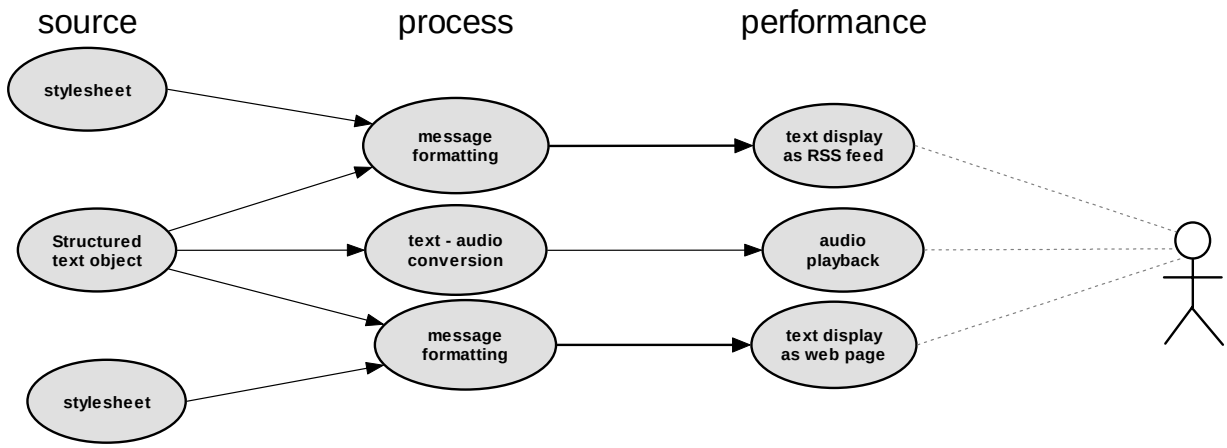


Figure 6: An illustration of the process required to interpret a structured text document

A structured text document is composed of mark-up that encapsulates fragments of text. Through the use of certain tags, the creator is able to specify the meaning of the text and how an interpreter should handle it. In isolation, the text and semantic markup located in an XML document contains the Information Content to be preserved. However, it does not indicate the method in which it has been, or should be, presented to the user. In order to record details of the performance, the digital archive must describe the rendering method that has been used and the relationship structure that is visually established.

For analysis purposes, the InSPECT Project has adapted the Digital Preservation Testbed classification scheme and attempted to categorise each property into the five groups, as illustrated in section 2.3 above. Several problems were encountered, notably the difficulty in classifying properties to a specific category. It was also recognized that the importance of certain properties was relative to the performance method. For instance, presentation formats such as HTML may contain a diverse set of structured and unstructured information that possess complex, and often poorly defined inter-relationships. The project identified 20+ entities that must be specified in order to describe the significant properties of presentation formats (figure 7).

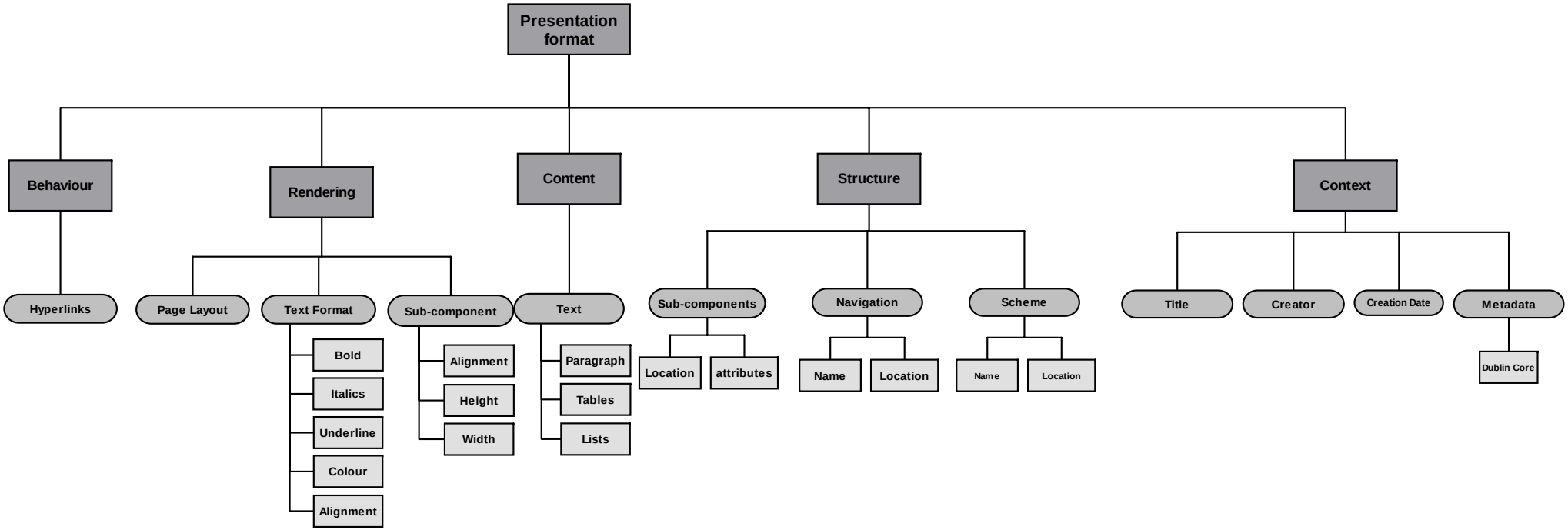


Figure 7: Significant properties of a structured text

Content
The content is the primary Component of a structured text document that must be preserved. For structured text documents, the characters (words, numbers and punctuation) contained in the document that are presented to the user onscreen, e.g. the lines of a play script, paragraphs in a document, etc. must be preserved. A simple method of identifying the success of text conversion is to measure the number of characters that are contained in mark-up tags. However, the number of characters that appear in the raw text is likely to differ from the number of characters that are displayed on screen. Several reasons may be identified for the discrepancy, including the use of tags to alter the appearance of the text (i.e. bold, underline, etc.) and the use of specific syntax to indicate characters declared illegal in certain character encodings, e.g. the use of " to indicate quotation marks in ISO 8859[r]. A potential solution may be to standardise the output of the Content, prior to its measurement (See Measurement Challenges).

Context
'Context' descriptions should establish the provenance of the Record's creation e.g. the title, creator, email, creation date and other pertinent details. It may also include details of the public or private standards used for the creation of the resource, e.g. TEI, DDI guidelines. Some context information may also be required, to indicate the method by which a value should be interpreted. For example, a list of values may contain conditions indicating that a scaled list containing numeric values is measured in inches.

Structure
The semantic and/or presentation structure is essential for the correct interpretation and rendering of the Record. The classification of structural entities is particularly difficult due to the likelihood that structural information will also perform other functions, for example provide context information. Several types of structural relationship may be identified, according to the type of structured text:

- Semantic tags defined in the structured text file that indicate the relationship between entities (e.g. 'director' may be a sub-element of 'performance');
- The relationship between visual Components e.g. logical groups of information, such as paragraphs, tables, lists, indexes, etc., that must be rendered.
- Parent-child relationships between the primary (text) and secondary Components (e.g. images, audio, video, schema, etc.);
- Sibling relationships with similar Records referenced in the text document, e.g. hyperlinks to other web pages.

Rendering
The rendering of structured text refers to the re-creation of its appearance to the user. It should be interpreted as being applied to the 'performance' of the text document in a particular environment, such as a web browser, and is not intended to describe every method of rendering that may be achieved. Therefore, it is possible to consider various scenarios that have different requirements

- *Web Feed* - A 'web feed' or 'syndicated feed' is a data format used to distribute frequently updated content to users. Examples of web feed formats that are in common use include the Atom and RSS standards. For web feeds, the following aspects are considered to be essential.
- *Web Page* – A data format that is suitable for distribution on the World Wide Web and may be accessed through a web browser.

The following properties may influence the rendering and, potentially, the implied meaning of the text:

- Text formatting – Formatting that alters the visual appearance of words to distinguish them from other elements on the page. The options available for formatting text differ between markup languages. Common examples include the use of bold, italics and underline, colour, and font size.

- Page Layout – The layout of elements on the page may have some significance, as defined by the structure e.g. text boxes must be placed in sequential order.
- Location of Components – The location of secondary Components, such as images on the page is likely to have some significance. Details of the height, width and alignment will be beneficial.

The relationship between the implied meaning and rendering of text is poorly defined and used inconsistently. For the purpose of preservation, it is not necessary to recreate the exact appearance of the text. However, it is preferential to maintain some formatting and visual relationship between page elements.

Behaviour
'Behaviour', in the context of structured text, refers to events that may be executed by an agent, a user or a software system that alters some aspect of the Record. Common examples of interactivity that may be contained or referenced by a text page include scripts that show the current date, customize the display to the user's browser or tailor information to user requirements (e.g. weather reports), as well as navigation between pages. The navigation structure is the single Behaviour that may contribute to the re-creation of the performance and should be maintained in subsequent manifestations. Three types of navigation may be identified:

1) Intra-Record navigation (internal) – Navigation between elements in the page.
2) Inter-Record navigation (internal) – Navigation between resources located internal to the collection.
3) Inter-Record navigation (external) – Navigation to resources located on external sites

Behaviours that alter the operation of the page, in the context of structured text documents is a complex area that requires consideration of many different methods with which a user may interact with a digital resource. Further tools development in the area of Transactional Archiving[S] is required to identify significant properties of user interactivity.

3.1.3. Measurement Challenges

The identification and recording of the characters and markup in the Record itself is an effective language-independent method of measuring the significant properties of a digital Record. However, two problems may be identified that limit the assessor's ability to gain a detailed understanding of the property:

1. *Malformed tags* - Malformed tags are one of many common errors found in structural text, particularly HTML files, that may affect the assessor's ability to measure the document structure. The term refers to the incorrect expression of opening or closing tags in a file, e.g. an opening paragraph tag is defined, but the closing tag is missing, or tags are improperly nested (e.g.). This may present problems when attempting to record the document structure.

2. *Special characters* – Many character encodings and markup languages reserve certain characters for use in particular circumstances and specify that any other use in a text document is prohibited. Common examples include left (<) and right (>) brackets, ampersands (&) that are used for the definition of HTML elements. However, there is often an alternative method of expressing the character that can be rendered, e.g. < for left bracket, & for ampersand, etc. Although the representation of such characters is not an issue, they present problems if the digital archive is measuring the success of a file conversion by counting the number of characters contained in the Record.

The value of measurements extracted from structured text in their submitted format may be questioned if it is likely that the Record is affected by the issues identified. A software application may misinterpret the relational structure of the document, or miscount the characters. The digital archive may be required to normalize the content prior to the creation

of a canonical list and the measurement of the Record properties. Software code[5] exists to correct the majority of malformed tags. However, the process is automated and may change the rendering of certain characteristics. Similarly, special characters may be normalized to reduce the likelihood that anomalies will occur. The W3C has developed the 'Canonical XML' standard that may serve as a method to reduce the complexity of a Record, by reformatting text content. By normalizing an XML document, the encoding method is changed, white space is removed, default attribute values are added, special characters are reformatted to system-legal characters, and comments are stripped.

A second method of recording the significant properties of rendered text is to measure the visual relationship between the Components.

3.3.4. Significant properties of structured text

A list of the significant properties of structured text is forthcoming.

[5] One possible example is the open source tool, HTML Tidy

3.2. Email

3.2.1. Definition

Electronic mail, commonly shortened to email, is a method of transmitting messages over an electronic communication system[t], as opposed to any distinction between content types. The specification for email messages is defined in several documents, collectively called the Multipurpose Internet Mail Extensions (MIME). The specifications indicate that an email must consist of two Components:

1. *Header* – Structured data that provides information about the sender (name, e-mail address), the path that was taken to deliver the message, the intended recipient (name, e-mail address), an indication of the subject and other relevant information.
2. *Body* – Unstructured text that, in the majority of cases indicates the primary content of the message.

The information encoded in the header is well defined and relatively consistent. The message body, in contrast, may contain diverse types of unstructured content, as specified by the creator. An email may encapsulate a diverse range of content types, including text, still images, audio, moving images, interactive resources, dynamic scripts, and other information. The communication of the Performance may be unique for each email, composed through interaction between disparate Components. For example, an image may be displayed in the body of an email message, or displayed as an attachment.

In practice, email may be instantiated as a compound object that contains a diverse set of structured and unstructured information that possesses complex inter-relationships. To some extent, an email may be considered an application of structured text. The description and provenance information contained in the header is well defined, complying with the appropriate MIME specification. The content of the message, contained in the body of the email is less defined, structured according to the creator's requirements. Each component – still image, moving image, sound - may possess significant properties that must be considered to preserve the message contained in the file. For analysis purposes, the report will focus on the properties associated with an email that primarily contains textual content and the relationship to other secondary Components. The ability to understand the performance contained in secondary Components, which are expressed using diverse methods and encoded in many different formats, is considered to be outside the scope of analysis. However, it is recognized that email attachments and other Components may contain information to support the Information Content found in the email itself.

In a digital environment, authenticity may have two different interpretations: an archivist is likely to consider authenticity in terms of the provenance of the Information Content by examining evidence of the creation process; a digital librarian is likely to consider authenticity in terms of the re-creation of the Information Content. Both definitions of authenticity are valid when applied to emails, which may contain information that may be interpreted differently, according to who sent it. To clarify, the InSPECT Project is concerned with the latter approach, regarding the re-creation of the Information Content. However, several elements of the former may also be relevant. The characteristics of an email considered essential to its performance are those that contribute to the re-creation of the Information Content and establish, to some extent, its origin. The integrity requirements of an email apply to the following aspects:

1) Successful re-creation of the message body
2) Establishment of the provenance of the message
3) Identification and re-creation of relationships between Record Components

4.2.2. Application of the Performance model

For the successful performance of an email, a digital archive must reproduce the Information Content located in the message body, a description of the message's provenance and the

relationship with secondary Components (email attachments and other embedded Components). Figure 8 provides a simple diagram to illustrate the performance of an email. In a performance metaphor, the Creator/Sender is the character in the play who has been assigned the message body as dialogue, to be processed and communicated through a performance. The message may be interpreted and expressed using different software – it may be communicated as text through email applications or text editors, interpreted by a text-to-audio converter and read to the recipientor various other methods may be used[u].

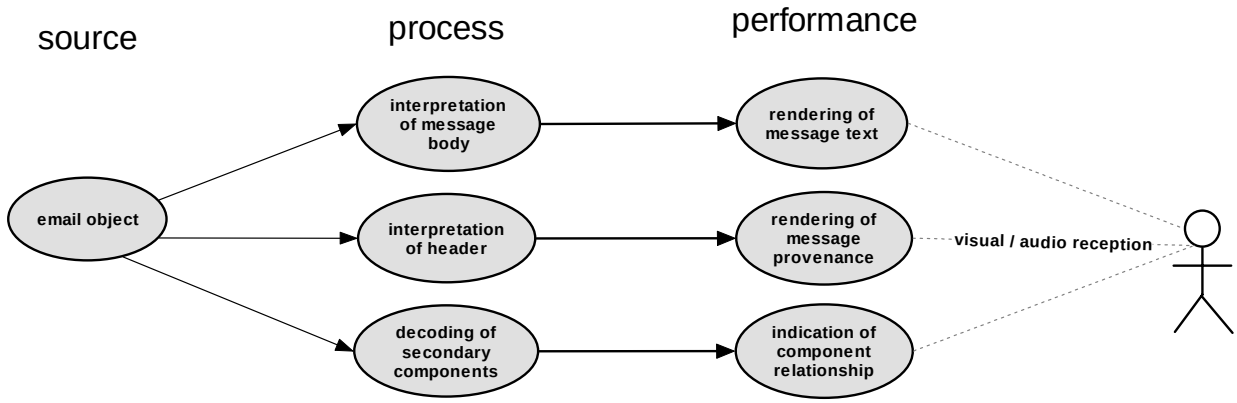


Figure 8. An illustration of the process required to interpret email data and re-interpret it for the user.

The significant properties identified by the InSPECT Project team are broadly similar to those identified in the Digital Preservation Testbed. In total, the project team identified 28 properties of various types that must be preserved (figure 9). However, there are minor differences in the recommended measurement method and the classification of certain entities as optional.

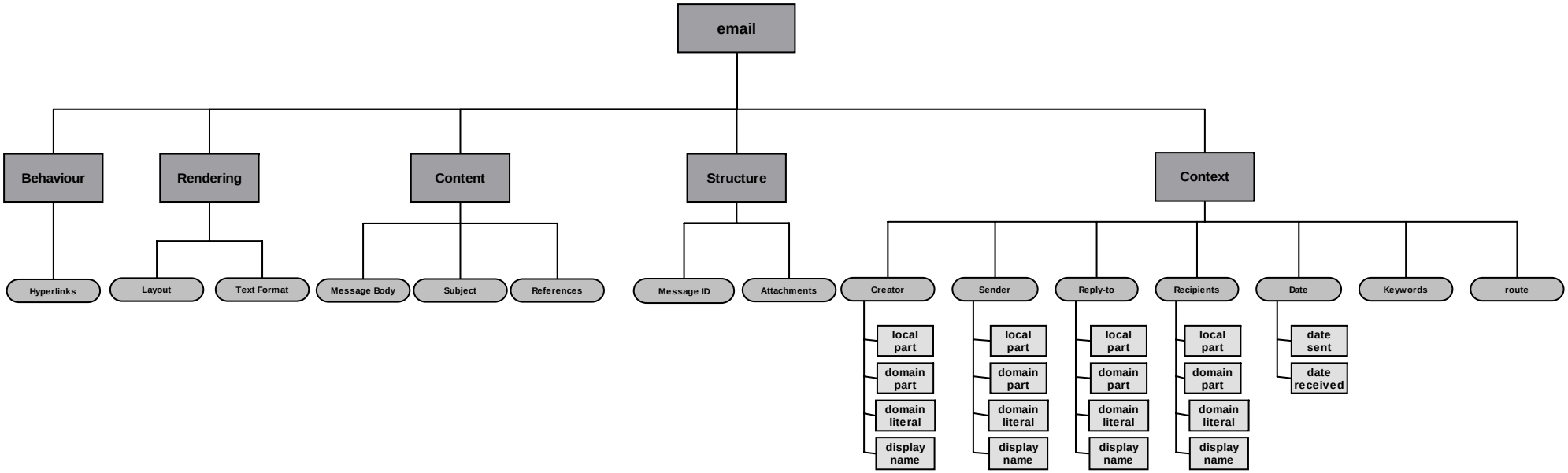


Figure 9: significant properties contained in an email

Content
In most circumstances, the message body is the primary type of content that must be preserved[6]. The email body may contain unstructured or semi-structured information that appears as text paragraphs, tables, lists and other categorization methods. Information may be encoded as plain text, HTML, or Microsoft Rich Text format. These three allow the creator to specify attributes of the appearance, such as layout, colour, size, etc. The Subject line is a second type of content that is often used to summarise the purpose of the email (e.g. 'meeting date'), its relation to previous messages[7], or to communicate other types of information. A simple method of measuring that the two content types have been stored and reproduced correctly is to count the number of characters in the email. However, mismatches may occur when comparing the character count as a result of the use of special characters (see. 3.1.3 for a further explanation).

Context
The provenance of an email may be established through the recording of three broad categories:

- *Origin and intended destination* – The identification of those responsible for the creation of the email and its intended recipient is a key Component in establishing the message's authenticity. The recipient may be indicated as the primary, secondary, or tertiary recipient through the use of the To, Carbon Copy (CC) and Blind Carbon Copy fields[v]. An email header often contains brief details of the sender[8], indicating their name, email address and, in some circumstances, their IP address. The extent of information differs between emails, dependent on the stages of the lifecycle through which an email has progressed. For example, an email may be created by a director and forwarded by an employee or it may be created and distributed by a mailing list which removes originator information prior to receipt by the list software, etc. The origination date specifies the date and time at which the Sender indicated that the message was complete and suitable for delivery.[9].

- *Route and recipient* – A second, constituent Component to establish the authorship of the message is to identify the route that the email has taken to reach the recipient's mailbox. Emails contain detailed information on the servers through which they were passed, when being delivered to the recipient.

- *Purpose* – The purpose of an email may be identified by text in the header, indicating the subject of discussion, relevant keywords and other information.

Rendering
The rendering of an email refers to the visual layout of the email body. The appearance of the message content may vary between different software applications. The Information Content of the email must remain the same in different software applications. However, the performance of the content may differ, in terms of the font type, size, style, colour and formatting.

---

[6] A possible exception are emails used to transfer attached data that do not contain any text in the message body.
[7] When used in a message reply, the Subject field may begin with Re:, followed by the subject of the previous message.
[8] Scenarios in which the Creator and Sender differ include the use of mailing lists to send an e-mail, a company director writing an e-mail and sending it to a sub-ordinate for distribution, etc. If the Creator and Sender are identical and/or the originator is identified by a single mailbox, the Sender value is unimportant.

[9] The orig-date-creator value is dependent on the Sender's computer for time-settings, which may be accidentally or intentionally altered. It does not indicate the time that the message is transported by the delivery system.

The Digital Preservation Testbed analysis of email indicates that the appearance of the preserved email may change. However, the “original meaning of the digital record” (Digital Preservation Testbed, 2003, p27) must remain the same. For preservation purposes, it is recommended that some markup is maintained if it is used to imply meaning. However, the markup language (e.g. HTML, Rich Text) in which textual data is stored is unimportant.

The project recommends that the following information should be maintained:

- Layout – tables, lists
- Formatting – bold, italics, underline

In the layout of an email, information meaning may be lost if the distinction between table rows and columns is removed. Similarly, it should be evident when text is presented in a list. However, further details regarding the list type is considered to be of low importance. The text formatting is potentially controversial and many authors have argued that email should be a text-only medium. However, it is possible that a Creator may use text formatting for emphasis in an email, e.g. when critiquing work by another person that, if removed, would lose some of the meaning.

Behaviour
An email, similar to other types of structured text and word processing documents, may contain user-driven interactivity, such as hyperlinks for internal and external page navigation, dynamic advertisements (i.e. if it sent through a free webmail service or mailing list) and various types of scripts to provide customised content. The majority of behaviour may be considered a characteristic of the digital environment in which the message was created and superfluous to the Information Content itself.

The InSPECT Project has performed an analysis of possible behaviour that may accompany an email and recommends that one type - hyperlinks to content located on a local or remote storage facility - is converted to subsequent manifestations.

Structure
As a compound object, an email may contain several relationships. These may be sibling relationships between multiple Components (e.g. text and an image) that are essential for the interpretation of the Information Content, or parent-child relationships for supporting information, such as attachments. The structure of the email may be measured at two layers of granularity:

1) Identify the number of attachments that were provided with the email. An email attachment may perform several functions in regards to the understanding of the Information Content – it may contribute to the understanding, appearance, or composition of the Information Content, as defined by the Creator or Sender.
2) Identify the Record identifier and indicate the relationship between the email and the associated Records.

The Australian Government Email Metadata Standard (AGEMS)[w] may prove useful for describing the relational structure of e-mail.

3.2.3. Significant properties for email
Table 4 provides a preliminary list of the properties contained in an email that are considered to be contributing factors to maintaining its authenticity.

property value	component	property definition	function class	function description	significance level	constraint type [1]	constraint reason [1]	constraint unit [1]	constraint type [2]	constraint reason [2]	constraint unit [2]	datatype	comments
local - part	creator	The username or other identifier in use by the creator, prior to the @ symbol	context	creator	06	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				US - ASCII (RFC 2822) only, maximum. 64 characters (RFC 2821), case sensitive	
domain - part	creator	A host name or domain name that is used by a DNS to indicate the origin of the message	context	creator	06	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				US - ASCII (RFC 2822) only, maximum. 64 characters (RFC 2821), case sensitive	
domain - literal	creator	An indicator of the source domain of the message specified by its IP (numeric) address.	context	creator	02	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				[ ] . 0 - 9	The use of domain literals is discouraged in RFC 822.
display - name	creator	A plain text indication of the agent's name	context	creator	04	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				Alphanumeric	
local - part	sender	The username or other identifier in use by the creator, prior to the @ symbol	context	sender	10	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				US - ASCII (RFC 2822) only, maximum. 64 characters (RFC 2821), case sensitive	
domain - part	sender	A host name or domain name that is used by a DNS to indicate the origin of the message	context	sender	10	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				US - ASCII (RFC 2822) only, maximum. 64 characters (RFC 2821), case sensitive	
domain - literal	sender	An indicator of the source	context	sender	02	equality	Indicates the presence /	Boolean (present /				[ ] . 0 - 9	The use of domain literals is

property value	component	property definition	function class	function description	significance level	constraint type [1]	constraint reason [1]	constraint unit [1]	constraint type [2]	constraint reason [2]	constraint unit [2]	datatype	comments
		domain of the message specified by its IP (numeric) address.					absence of the value in the Record	absent)					discouraged in RFC 822.
display - name	sender	A plain text indication of the agent's name	context	sender	08	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				Alphanumeric	
local - part	reply - to	The username or other identifier in use by the creator, prior to the @ symbol	context	reply - to	02	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				US - ASCII (RFC 2822) only, maximum. 64 characters (RFC 2821), case sensitive	
domain - part	reply - to	A host name or domain name that is used by a DNS to indicate the origin of the message	context	reply - to	02	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				US - ASCII (RFC 2822) only, maximum. 64 characters (RFC 2821), case sensitive	
domain - literal	reply - to	An indicator of the source domain of the message specified by its IP (numeric) address.	context	reply - to	02	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				[ ] . 0 - 9	The use of domain literals is discouraged in RFC 822.
display - name	reply - to	A plain text indication of the agent's name	context	reply - to	02	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				Alphanumeric	
local - part	recipients - primary(No.)	The username or other identifier in use by the creator, prior to the @ symbol	context	primary Recipient	10	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				US - ASCII (RFC 2822) only, maximum. 64 characters (RFC 2821), case sensitive	

property value	component	property definition	function class	function description	significance level	constraint type [1]	constraint reason [1]	constraint unit [1]	constraint type [2]	constraint reason [2]	constraint unit [2]	datatype	comments
domain - part	recipients - primary(No.)	A host name or domain name that is used by a DNS to indicate the origin of the message	context	primary Recipient	10	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				US - ASCII (RFC 2822) only, maximum. 64 characters (RFC 2821), case sensitive	
domain - literal	recipients - primary(No.)	An indicator of the source domain of the message specified by its IP (numeric) address.	context	primary Recipient	05	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				[ ] . 0 - 9	The use of domain literals is discouraged in RFC 822.
display - name	recipients - primary(No.)	A plain text indication of the agent's name	context	primary Recipient	08	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				Alphanumeric	
local - part	recipients - secondary(No.)	The username or other identifier in use by the creator, prior to the @ symbol	context	secondary Recipient	10	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				US - ASCII (RFC 2822) only, maximum. 64 characters (RFC 2821), case sensitive	
domain - part	recipients - secondary(No.)	A host name or domain name that is used by a DNS to indicate the origin of the message	context	secondary Recipient	10	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				US - ASCII (RFC 2822) only, maximum. 64 characters (RFC 2821), case sensitive	
domain - literal	recipients - secondary(No.)	An indicator of the source domain of the message specified by its IP (numeric) address.	context	secondary Recipient	05	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				[ ] . 0 - 9	The use of domain literals is discouraged in RFC 822.
display - name	recipients - secondary(No.)	A plain text indication of the agent's name	context	secondary Recipient	08	equality	Indicates the presence / absence of	Boolean (present / absent)				Alphanumeric	

property value	component	property definition	function class	function description	significance level	constraint type [1]	constraint reason [1]	constraint unit [1]	constraint type [2]	constraint reason [2]	constraint unit [2]	datatype	comments
							the value in the Record						
local - part	recipients - other(No.)	The username or other identifier in use by the creator, prior to the @ symbol	context	other Recipient	10	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				US - ASCII (RFC 2822) only, maximum. 64 characters (RFC 2821), case sensitive	
domain - part	recipients - other(No.)	A host name or domain name that is used by a DNS to indicate the origin of the message	context	other Recipient	10	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				US - ASCII (RFC 2822) only, maximum. 64 characters (RFC 2821), case sensitive	
domain - literal	recipients - other(No.)	An indicator of the source domain of the message specified by its IP (numeric) address.	context	other Recipient	05	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				[ ] . 0 - 9	The use of domain literals is discouraged in RFC 822.
display - name	recipients - other(No.)	A plain text indication of the agent's name	context	other Recipient	08	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				Alphanumeric	
creation - date		The date and time that an e - mail was completed by a Creator	context	date	05	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				ISO 8601 (datetime)	
send - date		The date and time that an e - mail was completed by a Creator	context	date	10	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)				ISO 8601 (datetime)	
received - date		The date and time that an e - mail was received by a	context	date	10	equality	Indicates the presence / absence of the value in	Boolean (present / absent)				ISO 8601 (datetime)	

property value	component	property definition	function class	function description	significance level	constraint type [1]	constraint reason [1]	constraint unit [1]	constraint type [2]	constraint reason [2]	constraint unit [2]	datatype	comments
		recipient					the Record						
message - id		A unique, machine - processable identifier	structure	identifier	02	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)					
id - domain		An indicator of the domain in which the message - id is unique.	structure	identifier	02	equality							
message - id		A unique, machine - processable identifier	structure	reply - to - id	02	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)					
id - domain		An indicator of the domain in which the message - id is unique.	structure	reply - to - id	02	equality							
message - id		A unique, machine - processable identifier	structure	references	02	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)					
id - domain		An indicator of the domain in which the message - id is unique.	structure	references	02	equality							
subject		A short string that identifies the topic of the message.	content	subject	10	equality	Indicates the presence / absence of the value in the Record	Boolean (present / absent)	equality	Indicates the no. of characters	characterLength	ASCII. Maximum of 255 characters	
keywords		A list of important words and phrases that	context	keywords	08	equality	Indicates the presence / absence of	Boolean (present / absent)	equality	Indicates the no. of keywords	characterLength		

property value	component	property definition	function class	function description	significance level	constraint type [1]	constraint reason [1]	constraint unit [1]	constraint type [2]	constraint reason [2]	constraint unit [2]	datatype	comments
		might be useful for the recipient.					the value in the Record						
associatedComponents		An indicator that the message contained attachments or other associated components, in addition to the message body.	structure	relation	10	equality	Indicates the number of components that are associated with the Record	Integer					
hyperlink		An indicator that the message contains hyperlinks that must be maintained.	structure	hyperlink	08	equality	Indicates the presence / absence of hyperlinks in the Record that must be maintained	Boolean (present / absent)					
message - body		The message body of the Record that represents the primary content	content	message - body	10	equality	Indicates the number of characters contained in the message body of the Record	Integer					

Table 4: Significant properties of email and a preliminary indicators on measurement methods

3.3. Digital Audio

3.3.1. Definition

Sound in its original (analogue) state is a series of air vibrations (compressions and rarefactions), which are captured by our ears and then converted to electronic impulses for interpretation. Sound waves are commonly measured by their frequency and amplitude. The ability to hear sound is subject to a range of factors, including the receptive capabilities of the listener and the medium through which it is transmitted. Optimally, people can hear from 20Hz to 20000Hz (20kHz), although this decreases with age.

Digital audio refers to sound waves that are stored in an electronic format and subsequently reconverted into an analogue form, in order to be heard by the listener. Digital audio may be constructed within an audio manipulation package (born digital) or sampled from an analogue source and stored as a binary file. It may be encoded and stored using a number of methods - as a continuous waveform composed of samples taken at specific time intervals (e.g. Wave, MP3), as an instruction set that indicate the musical notes to be reproduced (e.g. Midi), multiple waveforms that are processed and reproduced in a non-sequential manner (e.g. Modules), and music notation (e.g. CMusic). Each encoding method possesses unique properties that must be preserved, in addition to the raw information of the waveform. In this report the analysis is limited to the investigation of digital audio encoded as a waveform. Waveform is considered an effective method to curate digital audio data[x] and a significant percentage of audio data stored in the AHDS and TNA digital archives are stored in this format.

3.3.2. Application of the Performance model

To perform an audio recording stored in digital format the audio file must be decoded by appropriate software and the resulting information processed by a digital-to-analogue converter. The result of the process is the playback of sound through one or more speakers as a series of air vibrations that the user is able to hear. The audio bit-stream is the primary component to be preserved in the digital record, accompanied by associated metadata to understand the provenance of the audio recording, e.g. creator, title of the work, or a transcript of the spoken information. Figure 10 provides a simple diagram to illustrate the performance of digital audio.

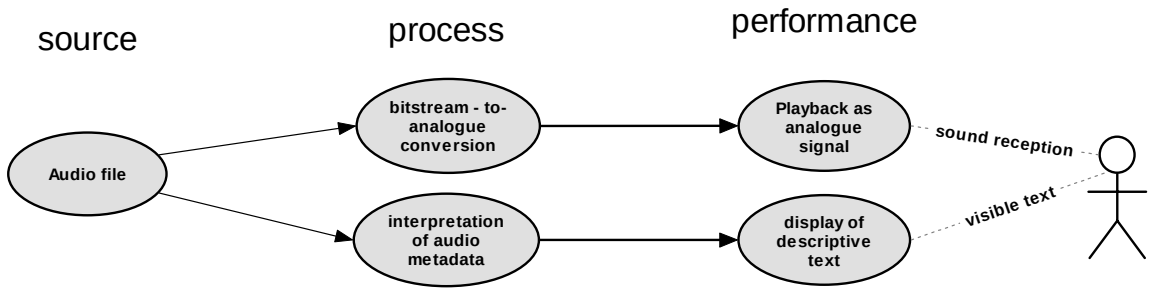


Figure 10. An illustration of the process required to interpret audio data and re-interpret it for the user.

The characteristics of a digital audio file considered essential to the performance are those that contribute to the successful re-creation of the audio recording. First, the integrity of the digital audio must be authenticated, by ensuring that the recording is complete (e.g. the recording length of a derivative matches an original) and has not been corrupted (e.g. sections of the audio stream have not been re-ordered). The quality level of the recording should also be maintained to some extent, to ascertain that the 'Information Content' contained in the audio continues to be understandable by the user. Secondly, the integrity of any description metadata that accompanies the audio file, either embedded in the file or stored separately, must also be maintained and converted, in order to maintain the context of the original recording.

Wilson et al (2006), in the JISC funded Moving Image and Sound Archiving Study, indicate that these requirements may be fulfilled by storing information on four properties that have some contribution to the accurate rendering of the digital audio:

Bit-Depth
Bit-depth indicates the number of bits used each second to represent the audio signal. It determines the dynamic range of recorded audio. For example 8-bit indicates that signal has been recorded using eight digits (e.g. 10010110); 16-bit indicates that the audio signal has been recorded using sixteen digits (e.g. 1001011011001010). A higher bit-depth will result in improved audio quality, but as a side-effect produces a larger file.

Sampling Rate
The sampling rate specifies the number of audio samples that are recorded per second. It is measured in Hertz (cycles per second). As a general rule, a greater number of samples may be recorded at higher sampling frequencies, i.e. the recording of audio at 44.1kHz or higher allows the recording of audio CD-quality data, while 8kHz produces telephone-quality audio.

Number of channels
A channel specifies the number of distinct outputs that may be used to playback sounds. An audio recording containing a single channel will output the audio through a single output; an audio recording that contains two or more channels may output samples to different outputs, as required.

Duration
Duration is the amount of time required to play the audio recording in full. It is a useful indicator to identify the loss of audio data, which may be caused by a misconfigured conversion process.

Metadata
For digital audio, metadata provides provenance information on the creator and the date of creation.

3.3.3. Classification of significant properties

For analysis purposes, the InSPECT Project has adapted the classification scheme for digital characteristics recommended by the Digital Preservation Testbed (see section 2.3 above). The classification scheme is composed of five broad categories: Content, Context, Structure, Appearance and Behaviour. The properties contained in a digital audio file are surprisingly easy to categorise into the five logical categories, unlike other file types previously discussed.

The audio and metadata bit-streams are located on the Component layer of the conceptual model described in the Methodology. After discussion, the project team classified the Content and Rendering categories as subsets of the audio bit-stream, the Context category as a subset of the metadata; and the Structure category as a combination of both Components. In total, the InSPECT Project identified eight properties in a digital audio file that must be identified and measured (figure 11).

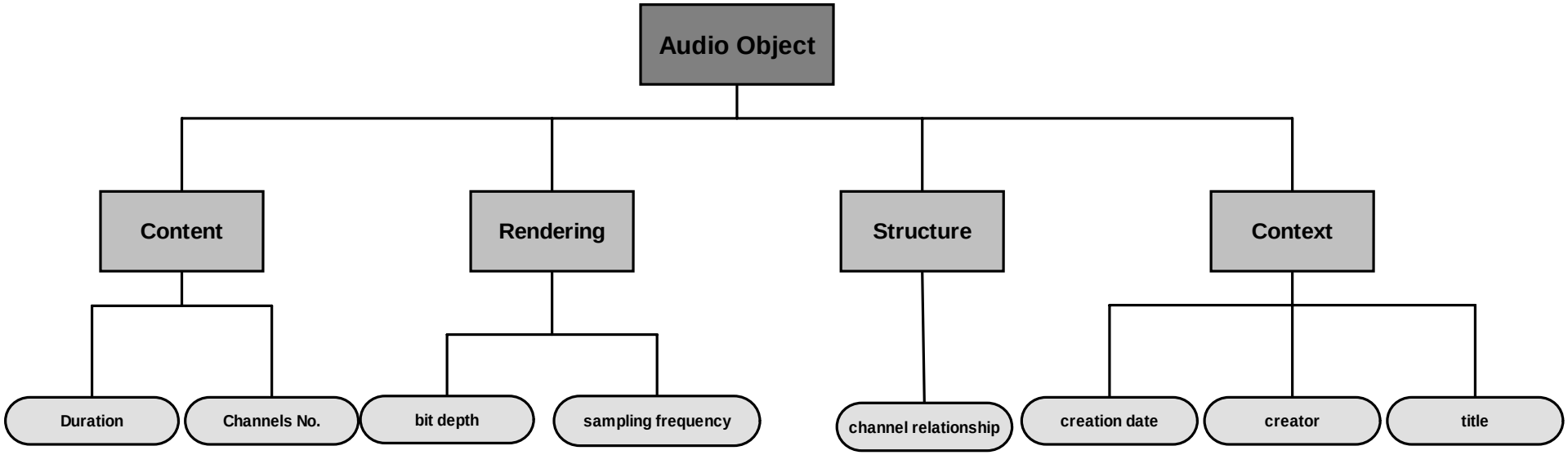


Figure 11: Significant properties contained in digital audio

Content
Digital audio contains sound that must be rendered to be understood. As an abstract entity, the significant properties of audio are the duration of the recording and the number of channels it contains. Both properties must be measured as positive integers, indicating the number of seconds required to replay the recorded audio and the number of distinct channels contained in the bitstream.

Rendering
The Content and Rendering categories are closely linked. The rendering of audio content for a performance is not required to be the same as the original performance. It may vary between different manifestations of the record. The preservation manifestation of the original recording should contain a bit depth and sampling frequency equivalent or higher than the value found in the original. However, a lower-quality derivative may be produced for distribution, e.g. internet streaming. Sampling Frequency is a positive integer value, measured in hertz. Common sampling rates for digital audio include 8000, 11025, 22050, 32000, 44100, 48000 and 96000. The bit depth specifies the amount of data contained in each sample, which has an effect on audio quality. It is measured as a one or two character numeric value, e.g. 8, 16, 24. Typically, a lower value denotes a poorer recording quality[10].

An institution may establish an Upper and Lower specification limit indicating the allowable deviation from the target value where a characteristic continues to be valid for the representation of the Information Content. For example, the 'target', or ideal sampling rate of a source may be 48,000Hz, while the tolerance specification defined by the archive indicates that a numeric value between 44100Hz and a hypothetical maximum of 96000Hz is acceptable for derivatives.

Context
Each file format differs in the type of metadata that may be stored. Broadcast Wave Format (BWF) contains five fields to identify the Creator, Creation date, Reference No, Description and Coding history[y]. The MP3 ID3 tags are designed for the classification of music, containing Title, Artist, Album, Genre and comment fields (Library of Congress, 2007a). A measurement of the metadata contained in a digital record will encompass the three criteria detailed in the Methodology section:

1) Identify if the digital record contains metadata
2) Review metadata fields and identify if they are populated.
3) Measure the property by recording information on the content contained in each field, e.g. by counting the number of characters, recording the values stored in each field.

The metadata associated with the recording provides provenance that may assist with the understanding of the Information Content, but is not essential to its access. However, subsequent validation may be performed to ascertain that it has been maintained.

Structure
The structural relationship between Components of an audio recording are rarely considered, except in circumstances in which the relationship is corrupted or lost. Two types of relational structure should be identified and maintained:

- The relationship between the audio stream and metadata to provide appropriate contextual information.
- The relationship between two or more audio channels which must be maintained to allow the correct rendering of the content. Each audio channel is directed to an appropriate speaker (left, right, etc.) if the relationship is maintained correctly.

[10] The bit rate – the amount of data transferred per second – is calculated by performing the following calculation: `Bit rate = (bit depth) x (sampling rate) x (number of channels).` For example, a record encoded using a 44.1 kHz sampling rate, 2 channels (stereo) and a 16 bit depth, the sample rate would be 1411200 bits per second.

These properties are particularly important if the digital archive operates a policy of storing digital information in its simplest form, by separating content and context into two or more files. It may not be possible to store all metadata values in the chosen normalization or distribution format, due to differences between format specifications. The second structural relationship may be important if each audio channel has been separated and stored in a different file.

Behaviour
No behavioural aspects of digital audio were identified that required conversion to subsequent manifestations.

3.3.4. Significant properties of digital audio

A preliminary list of properties considered to be significant for maintaining the authenticity of an audio record is provided in table 5.

property value	component	property definition	function classification	function description	significance level	constraint type [1]	constraint reason [1]	constraint unit [1]	constraint type [2]	constraint reason [2]	constraint unit [2]	datatype	comments
sampling-frequency	audio	A numeric value indicating the number of samples per second. The sample rate is measured in hertz (Hz) for audio recordings.	rendering		8	minimum / maximum	Indicates the minimum / maximum quality of the Record	Positive integer (hertz)					
bit-depth	audio	An indication of the quality of the recording, as indicated by the amount of data contained in each sample, measured by the number of bits. As a general rule, a bit depth of a low value denotes a poor quality recording.	rendering		8	minimum / maximum	Indicates the minimum / maximum quality of the Record	Positive integer (bits)					
Channels	audio	A numeric value that indicates the number of distinct channels that are part of the audio stream.	content / rendering		8	minimum / maximum	Indicates the number of values in the Record	Positive integer					
duration	audio	A concise indication of the length of the audio recording.	content		10	equality	Indicates the exact value of the property	positive integer (seconds)					
metadata		An indicator of the existence of metadata associated with the audio recording.	context		10	equality	Indicates the exact number of metadata elements in the Record	positive integer	equality	Indicates the no. of characters in the keyword	positive integer		

Table 5: Significant properties of digital audio and a preliminary indicators on measurement methods

3.4. Raster Images

3.4.1. Definition

Digital images may be encoded using a number of methods – the definition of geometric Components such as curves and polygons; a matrix of pixel elements that each contains colour information; and a combination of the two. Each encoding method possesses unique properties that must be preserved. For the purpose of analysis, this report investigates the significant properties of still images composed of a spatially mapped array of bits, or raster image. A raster image is composed of a rectangular array of pixels, that each represent a colour. The colour of each pixel may be defined by an RGB colour value, typically consisting of Red, Green and Blue values. Raster images are regarded as the most common type of image created and delivered over the Internet[z]. They are used for the creation and storage of many types of image, including photographs.

3.4.2. Application of the performance model

The performance of raster images requires the successful reproduction of the picture and metadata encoded in the digital record. The image is the core Component of the Information Content that must be preserved. It may be accompanied by metadata that indicates the provenance of the image's creation (e.g. person or institution responsible for its creation; date of creation) and a description of the Components contained in the image (e.g. a historic artifact, location that a photograph was taken, etc). Figure 12 provides a simple diagram to illustrate the performance of raster images.

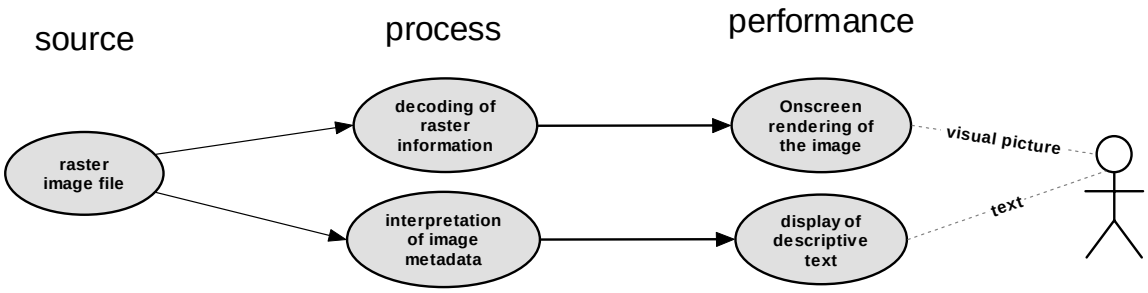


Figure 12. An illustration of the process required to interpret audio data and re-interpret it for the user.

3.4.3. Identification of significant properties

The characteristics of a still image file considered essential are those that contribute to the successful re-creation of the visual picture, as well as the associated metadata that establishes the context of the picture. The following properties are considered to be essential for understanding the image:

1. Resolution

The resolution is an indication of the number of pixels that are, or should be, contained in an image. The resolution is measured in pixels per inch (ppi) and has a deciding influence on the degree of detail that may be contained in an image. The amount of detail that may be stored in an image is proportionate to the number of pixels. For example, the scanning of an A4 document (9 x 12 inches) at 300 ppi will produce a digital image that is 2700 pixels x 3600 pixels (the dimensions of the original multiplied by the ppi); the scanning of the same A4 document at 600 ppi will produce a digital image that is 5400 pixels x 7200 pixels. The latter may therefore contain details that are not found in the smaller image. The scanning of a postage stamp (1 inch x 1 inch) at 300ppi will produce a digital image that is 300 pixels x 300 pixels. Although both items are scanned at 300ppi, they produce a different sized digital image. Therefore, it is often recommended that the pixel dimensions (pixels per inch) are used as an accurate method of referring to the size of a raster image[aa]. Details of the physical dimension of the image are useful, but not essential as the image may only ever exist as a digital, rather than physical, manifestation.

2. Bit Depth
The bit-depth refers to the amount of colour information held in relation to each individual pixel. A higher bit depth offers a greater number of available colours. A 2-colour image, often black and white, contains just 1-bit; a greyscale image typically contains 8-bits; and a full colour photograph typically contains 24-bits of information, offering 16,777,216 colours. The number of bits in an image has an effect on the file size, which increases significantly for 16-32 bit images.

3. Colour Space
The colour space of an image refers to the method of working with colours. It is influenced by the colour model in use – a mathematical formula that allows colours to be represented as tuples of numbers. Several colour models are available, including bitonal, grayscale, indexed colour, RGB and CMYK that are used for different types of images. The bitonal colour space uses two values, black and white; greyscale offers 256 shades between black and white; Indexed colours offer a limited palette of 216 colours which may be displayed on both Macintoshes and Windows PCs in a consistent manner. Computer monitors and televisions use RGB, to create colours as a combination of Red, Green and Blue colour values. It is common for designers to work with RGB and reduce the number of colours to Indexed colour for use on the World Wide Web.

3.3.4. Significant properties of raster images
A list of the significant properties of raster images is forthcoming.

Appendix A: Assessment Template

To gain an understanding of the significant properties associated with each data type, the following assessment template was developed by the project

Property title:
The title should provide an appropriate description of the purpose of the property. The title should be unique to avoid unnecessary confusion and, if possible remain consistent across similar file types. It should be relatively brief in its length. Examples: Sent Date, Receive Date, Subject

Property definition:
A formal statement that describes the purpose of the property.

Location
Indicate the intellectual Component to which the property applies. This may be left blank if it applies to the Record as a whole.

Function Classification
Classify the property according to each function indicated in the table. The purpose of each function is defined in the Assessment template report. The function description may be used to provide a more detailed description of the property function. One or more sub-categories may be entered for each category.

Function Classification	Function Description
Context	e.g. originator
Content	
Structure	
Rendering	
Behaviour	

Significance Level
Indicate the degree of importance that the property has to the creation of the Component or the Record. See section 2.2 for appropriate measurement values.

Property Constraints
Provide a description of the type of information that may be recorded for each value of a property. The assessor may add or subtract property tables as required.

Measurement 1

Property	Value
Unit	
Reason	
Type	
Value	
Type	
Value	

Measurement 2

Property	Value
Unit	
Reason	
Type	
Value	
Type	
Value	

Measurement 3

Property	Value
Unit	
Reason	
Type	
Value	
Type	
Value	

References

[a] Lynch, Clifford (1999). *Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information,* D-Lib Magazine 5:9. http://www.dlib.org/dlib/september99/09lynch.html
[b] CEDARS Project (2002). curl exemplars in digital archives. Retrieved on December 1, 2007 from: http://www.leeds.ac.uk/cedars/
[c] CAMILEON (2002). Creative Archiving at Michigan & Leeds: Emulating the Old on the New. Retrieved on December 1, 2007 from: http://www.si.umich.edu/CAMILEON/
[d] Bearman, D & Trant, T. Authenticity of Digital Resources: Towards a Statement of Requirements in the Research Process, *D-Lib Magazine*, June 1998. http://chnm.gmu.edu/digitalhistory/links/cached/introduction/link0.19c.digitalresourceauthenticity.html
[e] Wilson, A. (2007). Significant Properties Report. http://www.significantproperties.org.uk/documents/wp22_significant_properties.pdf
[f] IFLA Study Group (1998). Functional Requirements for Bibliographic Records. Retrieved on December 1, 2007 from: http://www.ifla.org/VII/s13/frbr/frbr.pdf
[g] Hunter, J & Lagoze, C. (2001). ABC Harmony Data Model. . Retrieved on December 1, 2007 from: http://www.metadata.net/harmony/ABCV2.htm
[h] OCLC & RLG (2005). Data Dictionary for Preservation Metadata. Retrieved on 15 February 2008 from: http://www.oclc.org/research/projects/pmwg/
[i] Anon (2007-09-10). PLANETS Project. Retrieved on February 4, 2008 from: http://www.planets-project.eu/
[j] ISO 15489 Information and documentation – Records management. Part 1 General.
[k] National Library of Australia. (nd). Guidelines for the Preservation of Digital Heritage. . Retrieved on December 1, 2007 from: http://unesdoc.unesco.org/images/0013/001300/130071e.pdf
[l] Rimington, J.D. & Trbojevic, V.M. (2000). Determination of ALARP in conditions of uncertainty. Retrieved on January 12, 2008 from: http://www.risk-support.co.uk/vmt_alarp_02.pdf
[m] Coyne. M et al (2007). The Significant Properties of Vector Images. Retrieved on February 15, 2008 from: http://www.jisc.ac.uk/whatwedo/programmes/programme_preservation/2008sigprops.aspx
[n] JISC: Significant Properties ITT (2007). Retrieved on February 15, 2008 from: http://www.jisc.ac.uk/fundingopportunities/funding_calls/2007/03/significant_properties_itt.aspx
[o] Rothenberg, J. & Bikson, T. (1999). Carrying Authentic, Understandable and Usable Digital Records Through Time. Retrieved on December 1, 2007 from: http://www.digitaleduurzaamheid.nl/bibliotheek/docs/final-report_4.pdf
[p] InterPARES (nd). InterPARES Project web site. . Retrieved on December 1, 2007 from: http://www.interpares.org/
[q] Digital Preservation Testbed (2003). From digital volatility to digital permanence: Preserving email. . Retrieved on December 1, 2007 from: http://www.digitaleduurzaamheid.nl/bibliotheek/docs/volatility-permanence-email-en.pdf
[r] Ramesh, M. (2000). iso8859-1 table. http://www.ramsch.org/martin/uni/fmi-hp/iso8859-1.html
[s] Pennock, M. & Patel, M. (2006). Digital Preservation Coalition Forum on Web Archiving. Ariadne, July 2006. Issue 48. Retrieved on March 17, 2008 from: http://www.ariadne.ac.uk/issue48/dpc-web-archiving-rpt/
[t] Anon (2008-01-13). E-mail. Retrieved on January 13, 2008 from: http://en.wikipedia.org/wiki/Email
[u] Digital Preservation Testbed (2003). From digital volatility to digital permanence: Preserving email. Retrieved on January 13, 2008 from: http://www.digitaleduurzaamheid.nl/bibliotheek/docs/volatility-permanence-email-en.pdf
[v] Resnick, P. (ed.) (2001). RFC 2822 (RFC2822). Retrieved on January 13, 2008 from: http://www.faqs.org/rfcs/rfc2822.html

[w] National Archives of Australia (2005). Australian Government Email Metadata Standard (AGEMS). Retrieved on January 13, 2008 from: http://www.naa.gov.au/Images/Email_Metadata_Standard_tcm2-911.pdf
[x] Knight, G. & McHugh, J. (2005). Preservation Handbook: Digital Audio. Retrieved on January 13, 2008 from: http://www.ahds.ac.uk/preservation/audio-preservation-handbook.pdf
[y] Chalmers, R. (1997). The Broadcast Wave Format - an introduction. Retrieved on January 13, 2008 from: http://www.ebu.ch/en/technical/trev/trev_274-chalmers.pdf
[z] Eadie, M. (2005) Preservation Handbook: Raster Images. Retrieved on January 13, 2008 from: http://www.ahds.ac.uk/preservation/Bitmap-preservation-handbook_d6.pdf
[aa] Anderson, S. et al (2006). Digital Images Archiving Study. Retrieved on January 13, 2008 from: http://www.ahds.ac.uk/about/projects/archiving-studies/index.htm